A Comparison of Clustering Algorithms for Use in Large Scale Gene Expression Analysis using Synthetic Microarray Data

C.E. Hart[*1], B. Bornstein[**1], T. Mann[**], J. Roden[**]
B.J. Wold[*], E. Mjolsness[**]

[*]*Division of Biology, California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125, USA*

[**]*Machine Learning Systems Group, Jet Propulsion Laboratories, California Institute of Technology Pasadena, 91109 CA USA*

We present a comparison of cluster of very computational complexities for use in large scale gene expression analysis. Synthetic data of the form of multi-dimensional hierarchical Gaussian trees were constructed and allowed us a direct comparison of clustering results to a 'ground truth". We introduce the use of using normalized mutual information (NMI) and receive operator characteristics (ROC) curves. Expectation maximization (EM), self organizing maps (SOM), k-means, and a phylogenetic clustering algorithms were compared varying the dimensionality, size, variance and structure of the synthetic data. SOM and EM performed nearly equally in low dimensions, but EM was able to continue promising performance

I. Introduction

The advent of large-scale gene expression analysis provides biologists with unprecedented amounts of quantitative data that comprise gene expression profiles of hundreds to tens of thousands of genes in many tissues, culture conditions, and genetic variants. The expression state of a cell can be assayed by extracting mRNA and then measuring the relative abundance each message. Currently, the two most prominent techniques for large-scale expression analysis involve 2-D DNA microarrays made by synthesizing complimentary DNA oligonucleotides on a glass slide via photolithography [1, 2] or by deposition printing of longer DNAs (usually cloned cDNA sequences prepared by PCR amplification) on a glass slide[3] (reviewed in [4]4).

A major goal in the analysis of large-scale gene expression data is to find sets of genes whose members have similar expression patterns. The challenge is to detect such similarities of expression pattern when working with sample numbers in the range from 10's to 1000's of different individual tissues, cell types, tumors, or growth conditions. Inspection and intuition, which have traditionally served

[1] Authors contributed equally

biologists well when they compared a few genes in 2 - 10 RNA samples, quickly fail when dealing with data sets of this scale. The application of clustering algorithms provides a way to reveal hidden structure in large datasets. A variety of clustering algorithms have been developed over the past three decades to deal with problems of this general form, and a few of these have recently been applied to large scale gene expression data [5-7]. The various algorithms make substantially different assumptions and operate by different mechanisms. This raises questions about how different algorithms will differ in detecting important structure in a given dataset, depending on properties that include both underlying biology and experimental noise. They will also likely differ in the nature and amount of artifactual association among genes and samples that they report. To best interpret the sensitivity and bias that each algorithm and parameter set introduces in a given application, it would be desirable to have a systematic method for evaluating the output of different algorithms and comparing them by a common metric.

In this work we compare three major types of clustering algorithms. The first type is phylogenetic clustering which uses a bottom up approach. The result is a very deep relatedness tree from which clusters can be extracted [7]. Expectation maximization (EM) algorithms [8], in contrast, work in a top down fashion when performed recursively or hierarchically ([5, 9]). Self organizing maps (SOMs) [6, 10] also use a top down strategy, but they differ because they first map the reults into low dimensional space which maintains proximity information from the higher dimensional space or gene expression trajectory.

There are also several choices of implementation that are expected to have impact on the output. The first of these is selection of the distance metric, a measure of relatedness, which can influence the sensitivity of an algorithm to particular features in the dataset. Euclidean and correlation distance metrics are two basic types whose implications we investigate. The Euclidean metric is sensitive to both magnitude and direction of change among data vectors. In contrast, a correlation distance metric is insensitive to the magnitude of change but retains sensitivity to the direction of change. Depending on the specific biological context and technical properties of the data, arguments favoring the application of either of these distance measures can be imagined, and examples are discussed. By using synthetic data that provides us with a known "ground truth" structure, we test the impact of both distance measures with data of differing structures.

A second potentially important implementation choice for some algorithms is the statistical model used. We explore here a collection of Gaussian distributions and discuss provisions for outlier rejection versus an alternative model that is a collection of Lorentsian distributions (the latter distributions have "heavier tails" and so give different weight to data vectors at relatively large distances from their cluster centers). We also discuss implications of distribution choice, depending on

whether all genes from an organism are assayed, as is typically done for yeast, compared with assays of smaller fractions of all genes, as is presently the case for mouse or human expression studies. This latter issue of complete versus incomplete gene sets is an especially important one for work done with any reduced complexity gene chip or any genome for which a fractional gene collection is all that is available.

A third potentially important choice for implementation of EM family algorithms is whether or not deterministic annealing is to be used. Deterministic annealing is a well established procedure designed to give superior global optimization by reducing the effect of local minima on the optimization process. Results from this study of synthetic data sets of varying structures identify candidate settings in which such local minima are likely to be problematic

In this paper, we introduce a framework for comparing and interpreting the results of many clustering algorithms. In addition to traditional mean/sigma plots and other visual inspection tools, we provide a mechanism to quantitatively assess cluster quality using receiver operator characteristic (ROC) curves [11]. We also introduce the idea of using a confusion matrix and normalize mutual information (NMI) scores [12] as a mechanism of interpreting the degree of agreement and disagreement between different clusterings of the same dataset.

Although the generation of these statistics aids in the interpretation of clustering results, it is not usually known what the complete "correct" clustering is for data sets in the current literature, since the underlying expression circuitry for even the simplest organism is far from fully understood. To address this, we employ synthetic data that has been constructed from several known architectures. This provides an additional way to evaluate the relative abilities of each of the clustering methods to find a known ground truth structure in the data one varies parameters such as the number of different measurements (dimensionality), relative separation of clusters from each other (variance ratio), and number of model genes for which data are provided.

I. Methods

Synthetic data with different known structures were generated and then used to assess the abilities and sensitivities of different clustering methods and the implications of choosing specific parameter sets. First, cluster centers were generated hierarchically using Gaussian distributions. A top level cluster was created with a variance of 1 and its center at the origin. Points were selected from this distribution to function as the cluster centers for the next level in the tree. The variance of these next level clusters is the product of the parent's variance and an adjustable parameter (the variance ratio). The variance ratio is held constant throughout the generation of a given synthetic data set and is equal to a child

cluster's variance over its parent's variance. The larger the variance ratio, the greater the cluster distributions overlap. Data points are eventually created from the bottom-most or leaf clusters. We choose to generate our synthetic data from a set of Gaussian distributions because of the frequency in which it occurs in nature (this is largely attributed to the law of large numbers) .

Using this general architecture or grammar we created four different types of synthetic data trees, two flat and two hierarchical. We generated Flat trees with either with 15 or 5 clusters ("15 trees" and "5 trees" respectively) whose centers were selected directly from the top level, or root, Gaussian. Hierarchical trees were generated with either 15 or 5 clusters whose centers were selected from the root Gaussian and then 3 more cluster centers were selected from each of those Gaussian (15,3 trees or 5,3 trees respectively). We adjusted the dimensionality of the dataset to 3, 10, or 30. For each dimension we also set the variance ratio to .1, .3, .5, or 1. Then, for each dimensionality / variance ratio pair we also varied the number of points created from 75, 750, 7500. In all, we created 36 different synthetic data sets.

Each of the above data sets were then subjected to the following clustering algorithms and the resulting cluster assignments were compared. This study is designed to compare: cross-validated expectation-maximization mixture of Gaussian (cv-EM-MoG) [13, 14] cv-EM mixture of Lorentzian (cv-EM-MoL), a deterministic annealing version of cv-EM-MoG [15] K-means [5], phylogenetic clustering (Xcluster) [7], and self organizing maps (SOM) [6, 10]. We also generated a "ground truth" clustering for each of the data sets which was derived directly from generation of the data.

The widely used expectation maximization (EM) family of clustering algorithms when performed recursively on a dataset provide a top down approach to partition the data into a set of clusters [8, 16]. Usually the clusters are assumed to be a multi-dimensional Gaussian distribution. Although, we also attempted the clustering assuming Lorentzian distributions. In either case, the underlying assumption is that each data vector was generated from one cluster and its value was obtained from a random sampling of its cluster's distribution. EM clustering algorithms attempt to discover the cluster membership for each data point by maximizing the likelihood. Normally using EM, the number of clusters must be preset. However, in the context of gene expression analysis it is unlikely the number of cluster is known. We used cross validation (cv) to estimate how many clusters should be used to describe our datasets [13, 14]. We performed this by maximizing the likelihood of each clustering as a function of the number of clusters.

We provide slightly more detail on the EM algorithms as used here [17]. We use EM with a diagonal covariance in the Gaussian, so that for each feature vector

component a (a combination of experimental condition and time point in a time course) and cluster α there is a standard deviation parameter $\sigma_{a\alpha}$. In preprocessing, each concentration data point is divided by its value at time zero and then a logarithm taken. The log ratios are clustered using EM. Optionally, each gene's entire feature vector may be normalized to unit length and the cluster centers likewise normalized during the iterative EM algorithm. This gives a variation of diagonal-covariance Gaussian mixture models which, for scalar variance, corresponds to a correlation distance metric rather than a Euclidean distance metric.

In order to choose the number of clusters, k, we use the cross-validation algorithm described by [13]. This involves computing the likelihood of each optimized fit on a test set and averaging over runs and over divisions of the data into training and test sets. Then, we can examine the likelihood as a function of k in order to choose k. Normally one would pick k so as to maximize cross-validated likelihood.

To initialize the EM clustering algorithm each cluster is given a mean and a variance. After initialization, the probabilities of each data point belonging to a cluster are calculated iteratively. Using these probabilities to weight cluster membership each cluster calculates a new mean and standard deviation using all the data points. This process is repeated until a local optimal solution is found. Cluster membership is then defined for each point as the cluster that they have the highest probability of belonging to. K-means is a computationally less intensive derivative of EM in which the algorithm has been modified to make cluster membership "hard", so that every data point at every step belongs only to the most probable cluster and has no influence on other clusters. Both EM and k-means are sensitive to the initial starting position of the clusters, for this reason every EM or K-means clustering was repeated 5 times varying only the random seed.

Phylogenetic clustering was the first cluster algorithms applied in the domain of large scale gene expression analysis [7]. It functions with bottom up strategy, where every gene begins belonging to a unique cluster. Each cluster is then compared to every other cluster and the two that are most similar to each other are combined and their mean is calculated. The process is repeated until only one "cluster" (the entire relatedness tree) remains.. Various approaches, including inspection, may then be used to select boundaries for membership in proposed discrete subclusters.

EM, k-means and phylogeneic clustering algorithms all rely on the calculation of distance between data points to determine the similarity of a data points or clusters. We performed all of the above clusterings using both a correlation distance

$$dist(x,\ y) = 1 - sim(x, y) = 1 - \left(\frac{\sum_{\dim} (x_i - \bar{x_i}) \cdot (y_i - \bar{y}_i)}{\sqrt{\sum_{\dim} (x_i - \bar{x_i})^2 \cdot \sum_{\dim} (y_i - \bar{y}_i)^2}} \right)$$

and a Euclidean distance

$$dist(x, y) = \sqrt{\sum_{\dim} (y^2 - x^2)}$$

metric.

Self organizing maps (SOM) [6, 10] attempt to map high dimensional data into clusters that exist in a much lower dimensional space (typically 1 or 2 dimensions). Each node, or cluster, in the low dimensional space represents some unique but general trajectory (or expression profile over a set of experiments) in the high dimensional space. Proximity in the low dimensional SOM corresponds to similarity in the high dimensional, experimental space. The algorithm is initialized by creating candidate clusters, or nodes, in the low dimensional space and then creating a mapping of each of the nodes into the higher dimension space. The relative position of the nodes in the low dimensional map is maintained in the higher dimensional space. Iteratively, a data point is selected at random and the node that is closest to that data point is moved towards it. The movements are large when the node is far from the data point and progressively as the distance to the data point decreases. The step sizes are also reduced as the number of interactions increase.

Twenty five independent clustering were run on 36 different data sets, to evaluate robustness, reproducibility, and ability to resolve the underlying data cluster membership and structure. Each clustering result was compared to the "ground truth" cluster structure and to other clustering runs of the same dataset by calculating receiver operator characteristic (ROC) scores [11] and normalized mutual information (NMI) scores [12]. NMI comparisons were also performed for results from both EM and K-means clustering runs that differed from each other only by the random seed used.

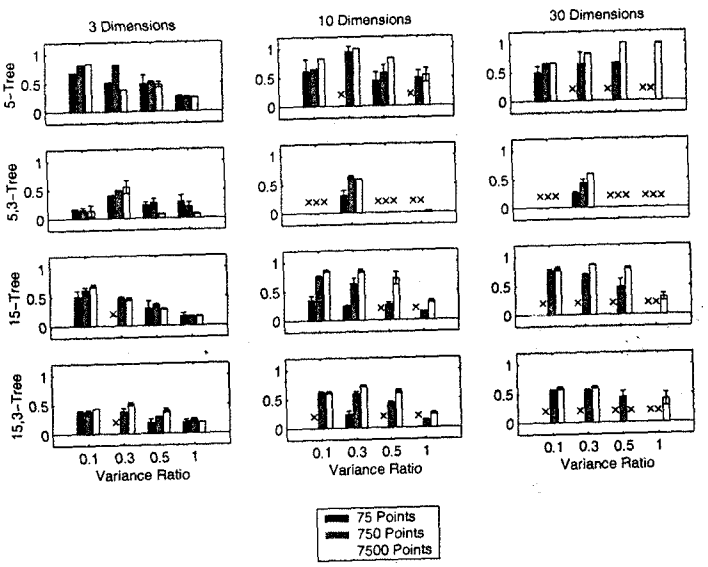
NMI generates a score based on the agreement between two clustering results (the score ranges from a value of 0 to 1). The average information contained within a clustering can be defined as $H(s) = \sum_{clusters} p \cdot \log_2 \cdot p$. Given two clusterings A and B and the average information shared between them equals H(A,B). The

information that one clustering relays about the other is equal to the mutual information I(A; B). From this NMI is defined as:

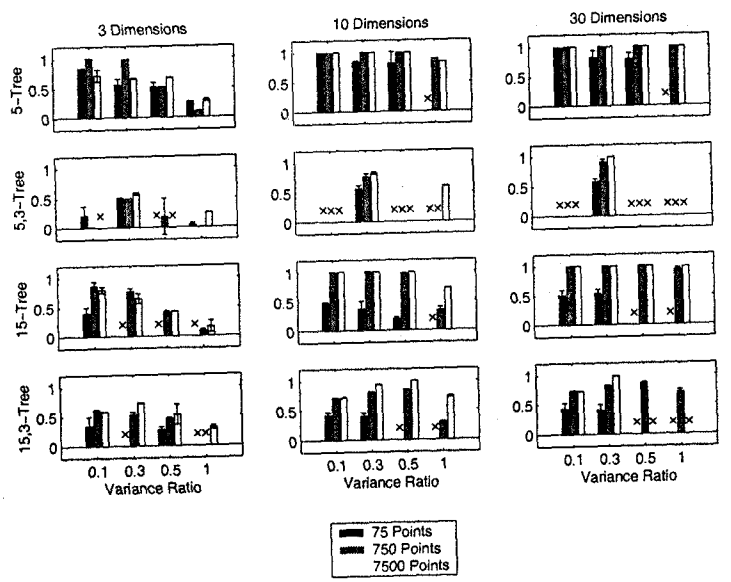
$$\eta(A) = \frac{I(A;B)}{H(A)} = \frac{H(A) - H(B) - H(A,B)}{H(A)} = 1 - \frac{H(A,B) - H(B)}{H(A)}$$

ROC analysis is a traditional technique used to evaluate fidelity of signal detection. An ROC curve is a plot of the proportion of true positives vs. false positive [11]. Here we have adapted it for "cluster gazing" by generating ROC curves in which the curve is a plot of the proportion of cluster members vs. non-members within a given cluster's most distant boundary. The area under this curve functions as good single value diagnostic measure of cluster overlap.

3. Results and Discussion

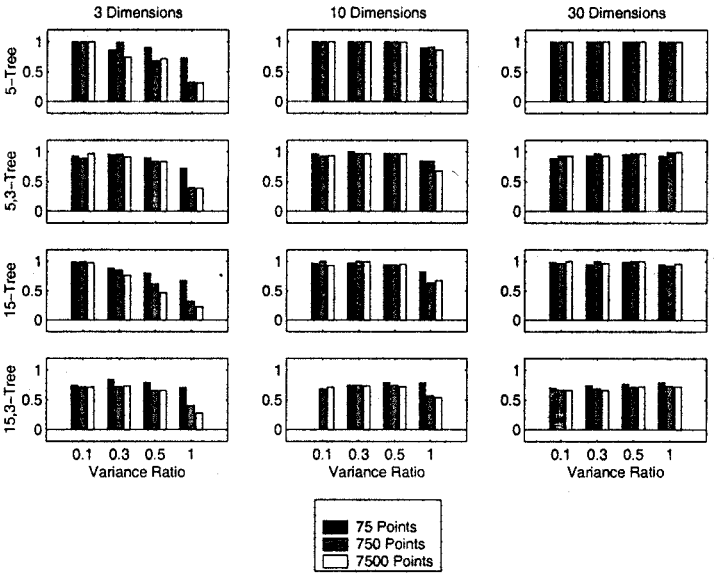


EM Correlation Summary



EM Euclidean Summary

Figure 1a:



SOM 1x25 Summary

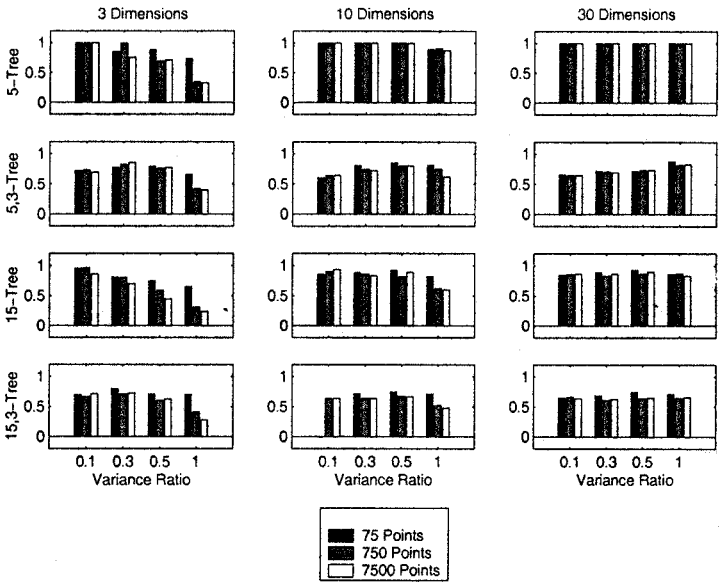


Figure 1b:

Summary of NMI statistics over varying synthetic data parameters. Mean NMI score plotted against variance ratio and number of data points. Black bars are 75 points, gray bars are 750 points, and white bars are 7500 points. Figure 1a) CV-EM MOG clustering with a Euclidian and correlation metric, Figure 1b) SOM clusterings using a map of either 1x25 or 5x5.

The NMI metric and ROC curves defined in Methods were used to measure and compare the effects on clustering performance that result from varying five parameters for implementations of three major types of clustering algorithms (EM, SOM, and phylogenetic). To permit us to compare output from each algorithm with a defined "ground truth" cluster structure, synthetic microarray gene expression datasets were generated as described above. Underlying this modeling exercise is a simple and plausible biological correlate for each cluster: a group of genes that are co-expressed under multiple conditions (dimensions in these synthetic datasets) because they share one or more functionally identical transcriptional enhancers or silencers (shared RNA turnover signals could also be at work, but they are conceptually identical to similar enhancers for the purposes of this study). Each of these hypothetical transcriptional enhancer or silencer types would cause the adjacent gene that it regulates to be transcribed similarly in response to a defined set of intra- and extracellular signals mediated by sequence specific DNA binding proteins and their associated coregulators.

Our Algorithm tests were performed on four different data structures: Two were "flat" cluster trees, one composed of 5 clusters and the other with 15 clusters. Each of these clusters would correspond biologically to a different group of genes, each defined by their use of functionally identical enhancers or silencers. The other two test cluster structures are hierarchical. The first consisted of five super clusters, each composed of three subclusters; and the second consisted of 15 super clusters, each containing three subclusters. In this dataset architecture, the biological correlate of each subcluster would again be one or more enhancers or silencers shared by members of the subcluster. However the hierarchical nature of the trees also indicates a relationship among subclusters of the same supercluster that makes them quite similar to each other. This might arise from any of several underlying biological models: A very straightforward one would be that each gene in the supercluster possesses one functionally similar enhancer and one novel one. The resulting expression patterns would share a common feature directed by the supercluster enhancer, and also have novel features corresponding to their unique enhancer. For each data structure, increasing dimensionality, model gene number, and variance ratios were tested.

Simple intuition might predict that clustering algorithms in general will perform best when clusters are well separated from each other (lower variance ratios) and when datasets are larger (here, gene number in each cluster). We found that this intuition held in general for all algorithms tested. However, each algorithm scaled differently (Table 1; Figure 1). Among the top-down strategies, EM/CV showed consistent improvements as the number of data points (model genes) increased. The EM/CV implementation that used a Euclidean distance measure was strongest when the datasets and dimensionality became large. This was in contrast to the Self Organizing Map which did not improve as a function of increasing number of

model genes. This difference is likely to be relevant to applications to current and future real datasets, as realistic gene numbers are between 6,000 (the gene number for complete yeast gene arrays or typical, reduced complexity human gene arrays) and 60,000 (human or mouse, estimated complete gene sets). Also, as predicted, more tightly clustered genes (low variance ratios of 0.1 for example) were generally easier for the algorithms to find correctly than "fuzzier" clusters with high variance. However, even this simple conclusion had interesting exceptions that were algorithm specific. Thus NMI scores for EM/CV Euclidean clustering runs were more successful at intermediate variance ratios (0.3 or 0.5), than at either 0.1 or 1.0 in the cases where the underlying data structure was hierarchical. Further investigation showed that at the lowest variance ratio, this algorithm was apparently readily able to correctly cluster the superclusters, but was unable to use increments small enough to find a define very narrow subclusters.

When using SOMs the structure of the low dimensional map is expected to have considerable impact on performance, and this was evident in our tests. We observed that for clustering our synthetic data 1x25 maps always out performed 5x5 maps and 1x10 maps always performed sub optimally. The extreme sub-optimal performance of the 1x10 maps on every dataset except the 5 member flat trees could simply reflect the fact that the starting node model doesn't provide enough elements to correctly partition the data. The 1x25 maps did much better for the larger cluster trees and for higher dimensionality. This may be reflect the fact that each cluster in the higher dimensional space is quite distant its neighbors. This might also explain the relatively poor performance of the 5x5 map, which would then be too limiting, forcing many more proximal similarities than the 1x25 map in the low dimensional representation of the clusters.

EM/CV and phylogenetic clustering (xcluster) performances both depended on the choice of distance metric. The Euclidean metric was uniformly superior to the correlation distance, but this was expected for this data architecture. Thus, we generated our synthetic data from a collection of Gaussian clusters, a Euclidean distance metric describes this space well, as it contains information on both direction and magnitude of each vector from the cluster mean, while the correlation metric ignores differences in magnitude. However we anticipate that in some biological settings the correlation metric will be superior, and the analysis here begins to define the penalty for its use. For example, a favorable biological setting for the correlation metric would occur when one wanted to uncover the regulatory similarity in two groups of genes of the following structure: Each member of group A has a weak basal promoter and each member of group B has a strong basal promoter, but genes in both groups are run by virtually identical transcriptional enhancers. Thus the direction of change with different stimuli would be governed by the enhancer and we would hope to infer this from membership in the same cluster. However, the magnitude of expression change

for A and B group members under differing conditions would be different due to strengths of their respective basal promoter types.

SOMs and EM performed similarly, with SOMs performing slightly better when the number of data points and dimensionality were low. As the dimensionality of the data set increased, and the dimensionality reduction by the SOM became correspondingly larger, the SOMs became less effective, at least within the range of node structures tested. EM on the other hand performed better when given high dimensions and a large number of data points. The same appeared true for phylogenetic clustering implemented by xcluster with agglomeration to set cluster boundaries.

Performance cost comparisons: The computational costs of the different algorithms are considerably different as are their expected further scaling properties at one log higher genes number and one to two log increases in dimensionality. These latter increases in matrix sizes are pertinent for anticipated studies with complete mammalian gene chips used with hundreds or thousand of different tumor samples, cell types or drug dose response courses. First, K-means is a relatively non-intensive implementation of EM, but it performed badly by both criteria used here under most data structures tested, and it was therefore not pursued in detail (data not shown). Among the others, SOMs are the least computationally sensitive and scales well. However, our analysis showed a trend disfavoring SOMs relative to x-clust-A (phylogenetic with agglomeration) at the highest dimensionality, gene number and variance ratio. A further investigation at still higher gene numbers and dimensionality seems warranted. The EM/CV algorithm is much more computationally intensive. A key question to be resolved in an extension of this study is the relative performance of each algorithm compared with its computational cost at the highest matrix sizes biologists are routinely likely to encounter.

Future Comparisons

Given the comparison framework presented here, further clustering algorithms can be systematically compared and tested on synthetic datasets exemplifying different biological assumptions. Prominent among the clustering algorithms yet to be fully examined in this way, is the xcluster algorithm of [7] (augmented with suitable "agglomeration" code to transform its binary cluster tree into other tree shapes), and the deterministic annealing and Lorentzian variants on the EM algorithm for mixture models.

Acknowledgments

This work was funded by LK Whittier foundation. We'd also like to thank the Wold Lab Group and Caltech Bioinformatics group for stimulating discussion.

References

1. **Lockhart, D.J., et al.,** ***Expression monitoring by hybridization to high-density oligonucleotide arrays [see comments].*** **Nat Biotechnol, 1996(13): p. 1675-80.**

2. **Fodor, S.P., et al.,** ***Light-directed,spatially addressable parallel chemical synthesis.*** **Science, 1991(4995): p. 767-73.**

3. **Schena, M., et al.,** ***Quantitative monitoring of gene expression patterns with acomplementary DNA microarray [see comments].*** **Science, 1995(5235): p. 467-70.**

4. ***The Chipping Forecast.*** **Nature Genetics, 2000(suppliment): p. 1-60.**

5. **Tavazoie, S., et al.,** ***Systematic determination of genetic network architecture [seecomments].*** **Nat Genet, 1999(3): p. 281-5.**

6. **Tamayo, P., et al.,** ***Interpreting patterns of gene expression with self-organizing maps:methods and application to hematopoietic differentiation.*** **Proc Natl Acad Sci U S A, 1999(6): p. 2907-12.**

7. **Eisen, M.B., et al.,** ***Cluster analysis and display of genome-wide expression patterns.*** **Proc Natl Acad Sci U S A, 1998(25): p. 14863-8.**

8. Dempster, A., NM Laird, and D. Rubin, *Maximum lielihood from incomplete data via the EM algorithm.* J.Royal Statistical Society,Series B, 1977: p. 1-38.

9. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proc Natl Acad Sci U S A, 1999. 96(12): p. 6745-50.

10. Kohonen, T., *Self-Organizing Maps.* Series in Information Sciences, 1995,second ed. 1997.

11. Swets, J.A., *Measuring the accuracy of diagnostic systems.* Science, 1988(4857): p. 1285-93.

12. FORBES, A.D., *CLASSIFICATION-ALGORITHM EVALUATION -5 PERFORMANCE-MEASURES BASED ON CONFUSION MATRICES.* JOURNAL OF CLINICAL MONITORING, 1995(3): p. 189--206.

13. Smyth, P., *Clustering using Monte Carlo Cross-Validation.* 1996.

14. Mjolsness, E., R. Castrano, and A. Gray, *Multi-Parent Clustering Algorithms for Large-Scale Gene Expression Analysis* . 1999.

15. Gold, S., A. Rangarajan, and E. Mjolsness, *Learning with preknowledge:Clustering with point and graph matching distance measures.* NEURAL COMPUTATION, 1996(4): p. 787--804.

16. Fukanaga, K., *Introduction to Statistical Pattern Recognition.* 1990.